

IET Radar, Sonar & Navigation

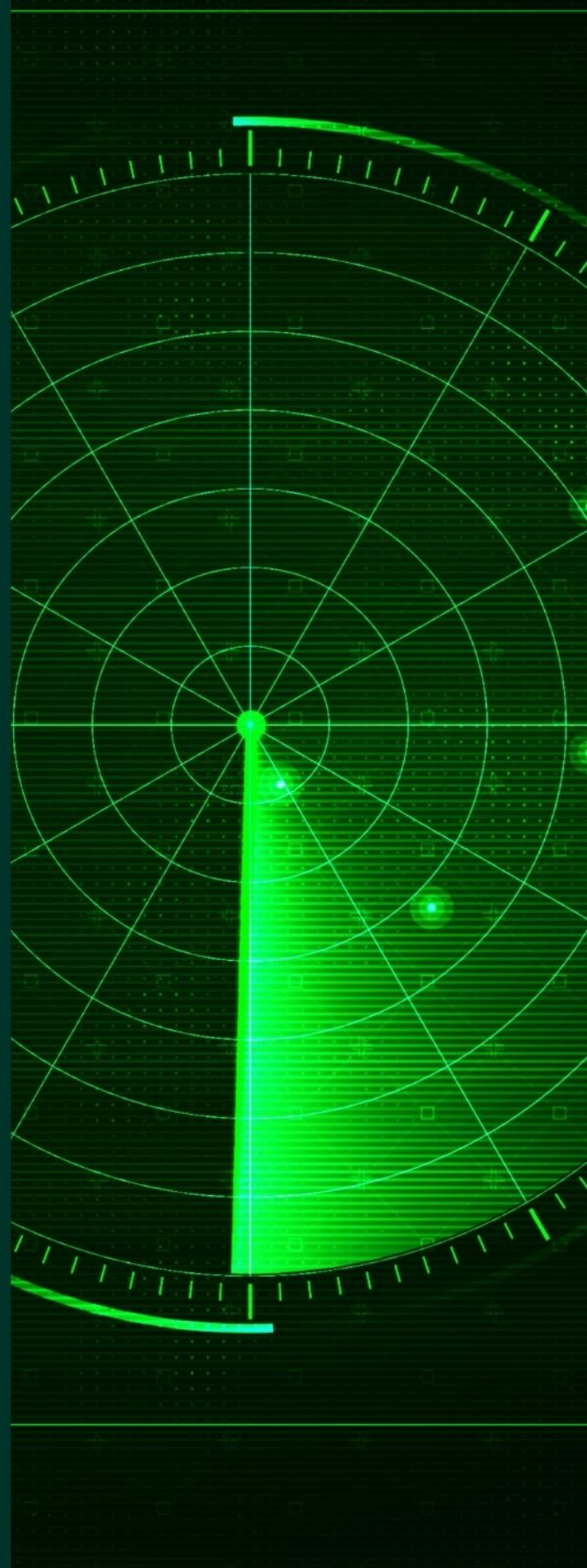
Special Issue Call for Papers

**Be Seen. Be Cited.
Submit your work to a new
IET special issue**

Connect with researchers and
experts in your field and
share knowledge.

Be part of the latest research
trends, faster.


[Read more](#)



The Institution of
Engineering and Technology

ORIGINAL RESEARCH

Multi-view convolutional neural network-based target classification in high-resolution automotive radar sensor

 Seunghoon Kwak | Hangyeol Kim | Gun Kim | Seongwook Lee 

 School of Electronics and Information Engineering,
College of Engineering, Korea Aerospace University,
Goyang-si, Gyeonggi-do, Korea
Correspondence
 Seongwook Lee, School of Electronics and
Information Engineering, College of Engineering,
Korea Aerospace University, 76, Hanggongdaehak-
ro, Deogyang-gu, Goyang-si, Gyeonggi-do, 10540,
Korea.

Email: swl90@kau.ac.kr

Funding information
 Ministry of Science and ICT, South Korea, Grant/
Award Number: 2021-0-00237
Abstract

In this study, a target classification method based on point cloud data in a high-resolution radar sensor is proposed. By using multiple antenna elements arranged in horizontal and vertical directions, pedestrians, cyclists and vehicles can be expressed as point cloud data in the three-dimensional (3D) space. To perform target classification using the spatial characteristics (i.e. length, height and width) of the target, the 3D point cloud data is orthogonally projected onto the xy , yz and zx planes, respectively, and three types of images are generated. Then, a multi-view convolutional neural network (CNN)-based target classifier using those three images as inputs is designed. To this end, a method for synthesising the detection results of three directions in series or in parallel is proposed. The proposed classifier can learn the spatial characteristics of the target by using the detection results of multiple viewpoints. Compared to the CNN-based classifier that uses only the detection result of a single plane as input, the proposed method shows 4.5%*p* higher classification accuracy in terms of the target with the lowest classification accuracy. In addition, the proposed multi-view CNN structure shows improved classification performance and shorter training time compared to the well-known deep learning methods for image classification.

1 | INTRODUCTION

In general, the most widely used waveform in automotive radar sensors is a frequency-modulated continuous wave (FMCW) [1]. In the FMCW radar system, the range resolution improves as the bandwidth used increases. Recently, as the frequency bandwidth available for the automotive radar sensor is widened, the range resolution is also improved. In addition, a multiple-input and multiple-output (MIMO) antenna system [2] has been adopted to ensure high angular resolution in a physically limited size. Therefore, unlike the conventional low-resolution radar systems, high-resolution radar systems that can express a single target as point cloud composed of several points have been developed [3–6].

Accordingly, with the development of a high-resolution radar system, a target classification method differentiated from a low-resolution radar system is required. In the low-

resolution radar system, the radar cross section [7] or range-Doppler characteristics of the target [8] were mainly used for target classification. However, in recent years, some studies have been conducted to detect targets and classify them in point cloud-based 77 GHz FMCW radar systems. For example, the authors in ref. [9] trained a kernel support vector machine by extracting several features from point cloud data to identify pedestrians and vehicles. In addition, a deep convolutional neural network (CNN) to classify multiple motions of a single person in a 77 GHz FMCW radar system was proposed in ref. [10], where the network was trained with accumulated point cloud data. Also, in ref. [11], the authors acquired point cloud data for several people using automotive radar and proposed a deep learning network to distinguish each person. Moreover, the authors of ref. [12] proposed the target classification network using self-attention mechanisms for millimetre-wave automotive radar systems. They also classified targets on the road by extracting

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *IET Radar, Sonar & Navigation* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

multidimensional feature vectors from the point cloud data and training the machine learning-based classifiers with those vectors [13].

In this article, we propose an effective target classification method for a high-resolution MIMO FMCW radar sensor. First, radar sensor data is acquired for representative targets, such as pedestrians, cyclists, sedans and sport utility vehicles (SUVs). The radar sensor we use has a range resolution of several centimetres. In addition, because the radar sensor uses antenna elements arranged in horizontal and vertical directions, the azimuth and elevation angles of the target can be estimated. Thus, using the estimated distance and angle information, the detected targets are expressed as point cloud data in a three-dimensional (3D) space. From the overall shape of the point cloud data, we can obtain information about the length, height and width of the target.

To effectively acquire spatial characteristics of the target, the point cloud data is orthogonally projected in three different directions (i.e. xy , yz and zx planes), and the detection results on three different planes are generated. Then, we design a CNN-based target classifier that uses three images as input. To this end, a method for combining the detection results of three directions in series or parallel is also proposed. Because the target detection result in each direction includes spatial information according to the type of the target, target classification accuracy can be improved by using the three images together. Finally, the classification accuracy of the proposed multi-view CNN-based classifier is compared with that of the classification method using only the detection result of a single direction. Moreover, the classification performance and training time of the proposed method are compared with those of the well-known deep learning-based image classifiers.

In summary, the main contributions of this study can be summarised as follows:

- Radar sensor data is acquired for four types of targets (e.g. pedestrians, cyclists, sedans and SUVs) from the high-resolution automotive radar sensor, and a point cloud-based target classification method is proposed.
- Unlike the conventional classification methods that use the target detection result in a single direction, the proposed method uses the detection results from multiple viewpoints of the target.
- To generate input data for training a CNN-based classifier, we propose a method for synthesising radar detection results of three directions in series or parallel.

The remainder of the paper is organised as follows. In Section 2, the fundamentals of target detection in the MIMO FMCW radar system are explained. Then, we describe the radar signal measurement environment and present the basic detection results in Section 3. Next, in Section 4, the multi-view CNN-based target classification is proposed and its classification performance is also evaluated. Finally, we conclude this paper in Section 5.

2 | TARGET DETECTION IN MIMO FMCW RADAR SENSOR

2.1 | Distance and velocity estimation using FMCW radar signal

The FMCW radar sensors are widely used in automotive radar systems, because they can estimate the distance to the target and the velocity of the target at the same time [14]. As shown in Figure 1, N_q chirps whose frequency increase linearly with time are sequentially transmitted in our FMCW radar system. In the figure, f_c , Δf and Δt represent the centre frequency, the operating bandwidth and the chirp duration of each chirp respectively. In addition, T_f represents the entire transmission period, and we define it as one frame.

If we assume that the transmitted FMCW radar signal is reflected from the k th target moving at a relative velocity of v_k at a distance of R_k , the received radar signal includes a time delay due to R_k between the radar and the k th target and a Doppler shift due to v_k between the radar and the k th target. Also, the Doppler shift can be expressed as $f_{d,k} = 2v_k f_c / c$, where c is the speed of light. Then, the received FMCW radar signal is down-converted to a baseband signal after passing through a frequency mixer and a low-pass filter (LPF). Finally, the baseband signal sampled at the analog-to-digital converter (ADC) can be expressed as

$$b[p, q] = \sum_{k=1}^{N_k} A_k \exp \left(j2\pi \left(\frac{2R_k \Delta f}{c \Delta t f_s} p + \frac{2v_k f_c \Delta t}{c} q + \frac{2R_k f_c}{c} \right) \right), \quad (1)$$

where N_k denotes the total number of targets and A_k denotes the amplitude of the baseband signal corresponding to the k th target. In Equation (1), p ($p = 1, 2, \dots, N_p$) is the index for time samples in each chirp and q ($q = 1, 2, \dots, N_q$) is the index for each chirp. In addition, f_s represents the sampling frequency and is the reciprocal of the sampling period T_s . A block diagram for the FMCW radar system is shown in Figure 2. The system consists of the waveform generator, voltage-controlled oscillator (VCO), amplifiers, transmit and receiving antenna elements (Tx and Rx),

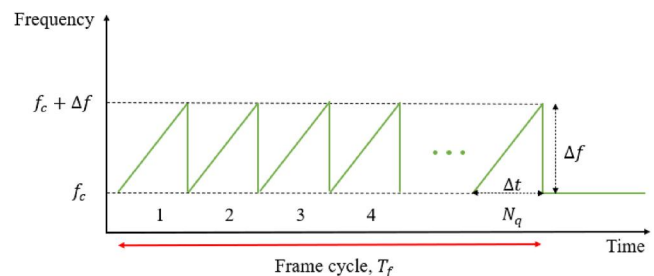


FIGURE 1 Signal transmitted from the frequency-modulated continuous wave (FMCW) radar system

frequency mixers, phase shifter, LPF, ADC and digital signal processor (DSP).

The time-sampled baseband signal of Equation (1) can be expressed in the form of a two-dimensional (2D) matrix, which is shown in Figure 3a. Then, the distance and the relative velocity to the target can be estimated by applying the Fourier transform in the direction of the sampling axis (i.e. p -axis) and in the direction of the chirp axis (i.e. q -axis), respectively, as shown in Figure 3. In summary, if the 2D Fourier transform is applied to Equation (1), the distance and velocity information of multiple targets can be extracted at once [15]. The 2D Fourier transform result of Equation (1) can be expressed as

$$B[r, s] = \frac{1}{N_p N_q} \sum_{p=1}^{N_p} \sum_{q=1}^{N_q} b[p, q] \times \exp\left(-j2\pi\left(\frac{pr}{N_p} + \frac{qs}{N_q}\right)\right). \quad (2)$$

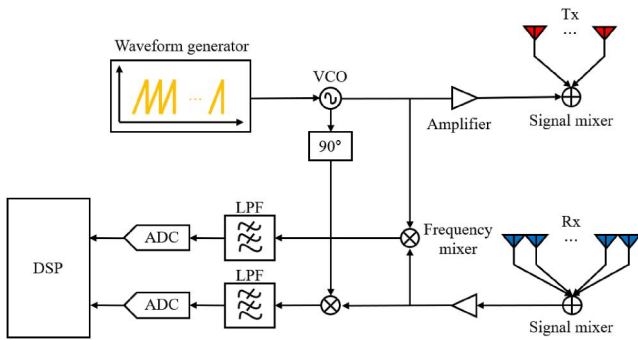


FIGURE 2 Block diagram of the multiple-input and multiple-output (MIMO) frequency-modulated continuous wave (FMCW) radar system

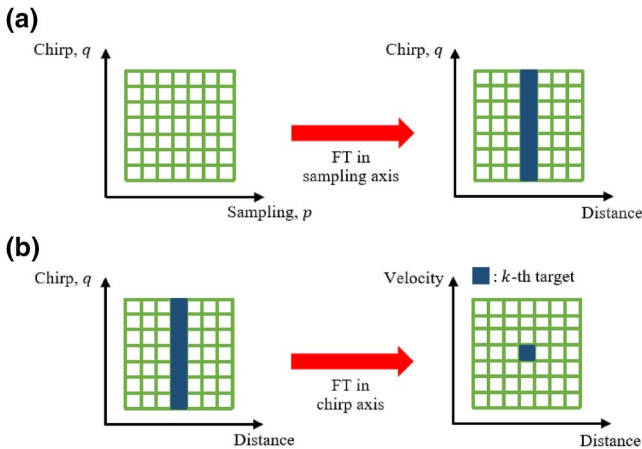


FIGURE 3 Distance and velocity estimation: (a) applying the Fourier transform in the direction of the sampling axis and then (b) in the direction of the chirp axis

2.2 | Angle estimation using MIMO antenna system

2.2.1 | Data cube generation for angle estimation

To estimate the angle information of targets, the MIMO antenna system consisting of multiple transmit and receiving antenna elements can be used. If we assume that the azimuth and elevation angles between the centre of the antenna and the k th target are θ_k and ϕ_k , respectively, Equation (1) can be expanded as

$$b[p, q, u, v] = \sum_{k=1}^K \alpha_k \exp\left(j2\pi\left(\frac{2R_k \Delta f}{c \Delta t f_s} p + \frac{2v_k f_c \Delta t}{c} q + \frac{f_c d_{t,a} \sin \theta_k}{c} (u-1) + \frac{f_c d_{r,a} \sin \theta_k}{c} (v-1) + \frac{f_c d_{t,e} \sin \phi_k}{c} (u-1) + \frac{f_c d_{r,e} \sin \phi_k}{c} (v-1) + \frac{2R_k f_c}{c}\right)\right), \quad (3)$$

where $d_{t,a}$ and $d_{t,e}$ are distances between transmit antenna elements arranged in azimuth and elevation directions respectively. Similarly, $d_{r,a}$ and $d_{r,e}$ are distances between receiving antenna elements arranged in azimuth and elevation directions. In addition, u ($u = 1, 2, \dots, N_T$) is the index of the transmit antenna element and v ($v = 1, 2, \dots, N_R$) is the index of the receiving antenna element respectively.

For example, in the MIMO antenna system where the number of transmit antenna elements is N_T and the number of receiving antenna elements is N_R , the number of receiving channels can be virtually increased to a maximum of $N_T \times N_R$ [2]. In general, because angular resolution is proportional to the number of receiving channels [16], high angular resolution can be achieved with a limited device size by using the MIMO antenna system. When a total of $N_T \times N_R$ receiving channels are generated, a total of $N_T \times N_R$ 2D Fourier transform results of Equation (2) are generated. In other words, a data cube of $N_p \times N_q \times (N_T \times N_R)$ is created, which is shown in Figure 4.

2.2.2 | Angle estimation using digital beamforming

First, to estimate the azimuth angle of the target, receiving channels arranged in the azimuth direction are selected from among all virtual receiving channels. If we assume that the number of channels in the azimuth direction is N_A , a data cube of size $N_p \times N_q \times N_A$ is generated as shown in Figure 5a. Then, if indices corresponding to the k th target in the 2D Fourier transform result of each channel are r_k and s_k , respectively, a total of N_A sampled values can be taken as

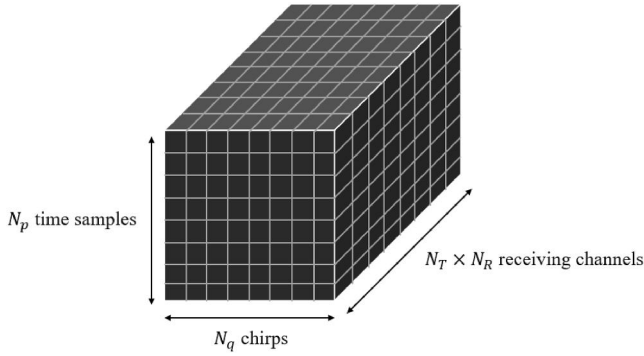


FIGURE 4 Generated data cube in the multiple-input and multiple-output (MIMO) frequency-modulated continuous wave (FMCW) radar system

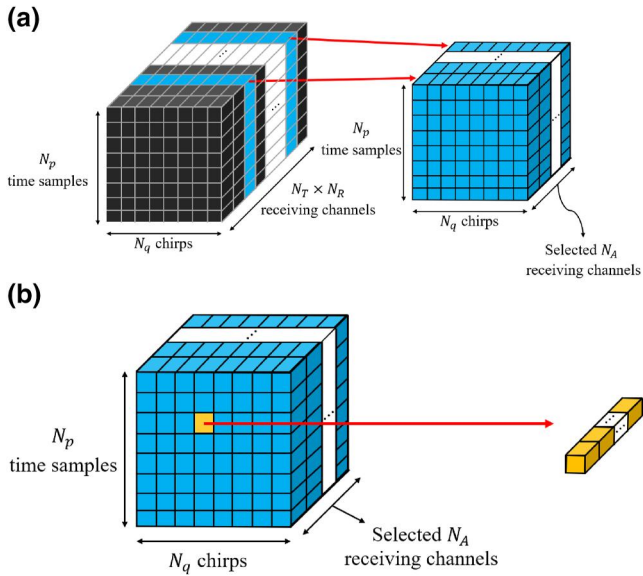


FIGURE 5 Receiving channel selection for azimuth angle estimation in the data cube

shown in Figure 5b. Thus, the signal vector composed of values sampled in N_A channels can be expressed

$$\mathbf{B}_A = \begin{bmatrix} B_1[r_k, s_k] \\ B_2[r_k, s_k] \\ \vdots \\ B_{N_A}[r_k, s_k] \end{bmatrix}. \quad (4)$$

Based on the signal vector of Equation (4), we use the conventional beamformer (i.e. Bartlett method [17]) for angle estimation, which is one of the digital beamforming techniques. To this end, we generate a correlation matrix using the extracted signal vector of Equation (4), which can be expressed as

$$\mathbf{R}_A = \frac{1}{N_A} \mathbf{B}_A \mathbf{B}_A^H. \quad (5)$$

In Equation (5), the symbol $(\cdot)^H$ denotes the Hermitian operator, so \mathbf{R}_A becomes a square matrix of size $N_A \times N_A$. Then, the normalised pseudospectrum of the conventional beamformer can be expressed as

$$P_A(\theta) = \frac{\mathbf{a}_A^H(\theta) \mathbf{R}_A \mathbf{a}_A(\theta)}{\max(\mathbf{a}_A^H(\theta) \mathbf{R}_A \mathbf{a}_A(\theta))}, \quad (6)$$

where $\mathbf{a}_A(\theta)$ is the steering vector considering the distance between the receiving channels in the azimuth direction. Finally, the azimuth angle of the target is determined by θ that maximises the value of the normalised pseudospectrum.

The process of estimating the elevation angle ϕ_k is similar to that of estimating the azimuth angle θ_k . The difference from the angle estimation in the azimuth direction is that N_E receiving channels are selected in the elevation direction. In addition, the distance between receiving channels considered in the steering vector $\mathbf{a}_E(\phi)$ is changed.

3 | ACQUISITION AND PROCESSING OF RADAR SENSOR DATA

3.1 | Measurement scenarios for acquiring radar sensor data

In our experiment, we used a radar sensor made by bitsensing Inc. (i.e. 79 GHz AIR 4D) [18], which is shown in Figure 6. This radar is a high-resolution imaging radar equipped with multi-chip cascading technology. A camera is also mounted on the radar, which can store images of the measurement environment. The detailed specifications of the radar sensor we used are summarised in Table 1. As shown in the table, the radar sensor uses 79 GHz as the centre frequency and 1.5 GHz as the bandwidth. In addition, the number of transmit antenna elements and the number of receiving antenna elements are 12 and 16, respectively. Because the antenna elements are placed in the horizontal and vertical directions, we can estimate the azimuth and elevation angles of the target. Then, the total number of chirps is 32 and 1024 time samples are obtained from each chirp. Also, the signal transmission period defined as one frame is 100 ms. The resolutions for range, azimuth angle and elevation angles are 10 cm, 2° and 5° respectively.

With the radar sensor, we acquired radar sensor data for various types of targets. The measurement subjects were selected as pedestrians, cyclists, sedans and SUVs, which are commonly seen while driving on the road. In addition, even within one target type, the measurements conducted several times while changing the subjects. The length, height and width information for the targets we measured is presented in Table 2. In addition, because the radar sensor is mounted on the bumper of the vehicle, the radar was installed at a height of 60 cm from the ground, as shown in Figure 7. Then, measurements were made at a total of 14 points, and the angle and distance between the radar and the target were different at each measurement point, as shown in Figure 8.



FIGURE 6 High-resolution radar sensor used in our experiment

TABLE 1 Specifications of the radar sensor

Parameter	Value
Centre frequency, f_c	79 GHz
Bandwidth, Δf	1500 MHz
The number of chirps, N_c	32
The number of time samples in each chirp, N_s	1024
The number of transmit antenna elements, N_t	12
The number of receiving antenna elements, N_r	16
Frame time, T_f	100 ms
Range resolution, ΔR	10 cm
Azimuth angle resolution, $\Delta\theta$	2°
Elevation angle resolution, $\Delta\phi$	5°

TABLE 2 Length, height and width information for each target class

Class	Length (mm)	Height (mm)	Width (mm)
Pedestrian	280–540	1680–1830	575–645
Cyclist	1680–1760	1750–1900	500–580
Sedan	4588–4995	1470–1485	1860–1870
SUV	4410–4750	1020–1715	1830–1930

3.2 | Target detection results in 3D space

By applying the radar signal processing technique described in Section 2 to the acquired radar sensor data, we can find out the distance to the target and the azimuth and elevation angles of the target, as shown in Figure 9. In general, because the distance estimated by the radar is the distance in the radial direction, the target information can be converted into (x, y, z) coordinates in the 3D space as follows:

$$\begin{aligned} x_k &= R_k \sin \theta_k \cos \phi_k, \\ y_k &= R_k \cos \theta_k \cos \phi_k, \text{ and} \\ z_k &= R_k \sin \phi_k. \end{aligned} \quad (7)$$

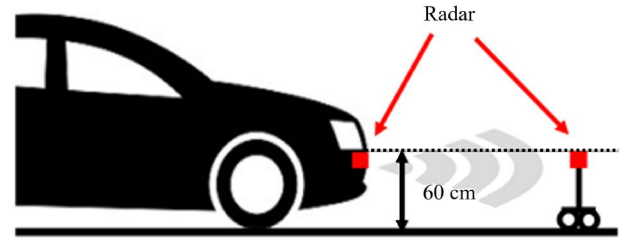


FIGURE 7 Mounting location of the radar sensor

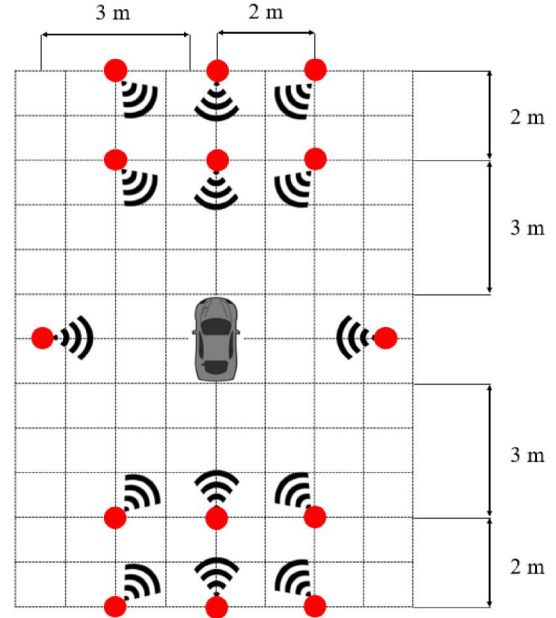


FIGURE 8 Measurement environment

For example, Figure 10 shows the detection result of processing the radar sensor data acquired 5 m behind the pedestrian. In Figure 10b, the points in the red box correspond to the points detected at the actual pedestrian location. In the FMCW radar system, the range resolution is determined as $\frac{c}{2\Delta f}$ [19], so it becomes 10 cm in our radar system. Thus, the pedestrian is detected as a point cloud composed of several points, as shown in the figure. The approximate shape and size of the pedestrian can be obtained from the point cloud data.

4 | PROPOSED TARGET CLASSIFICATION METHOD

4.1 | Multi-view CNN-based target classifier

4.1.1 | Preprocessing for input data generation

As mentioned in the previous section, the target information (i.e. R_k , θ_k and ϕ_k) was converted into points in a 3D xyz coordinate system. In general, because the CNN-based classifier uses an image as input, a process of making 3D point cloud data in the form of images is required. Therefore, we

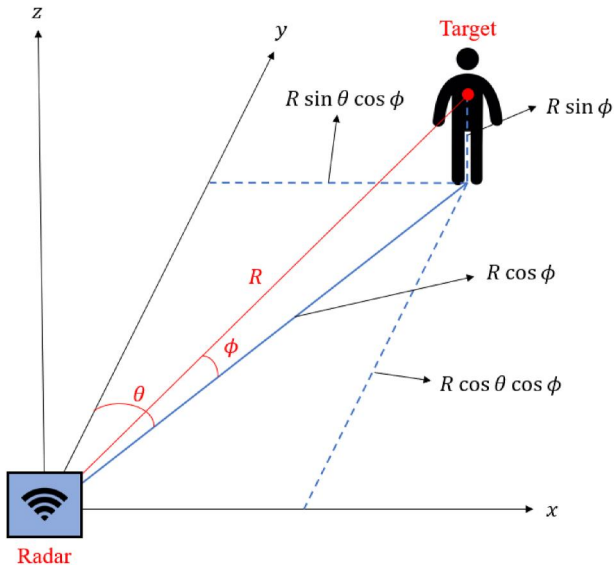


FIGURE 9 Target detection result expressed in xyz coordinate system

(a)



(b)

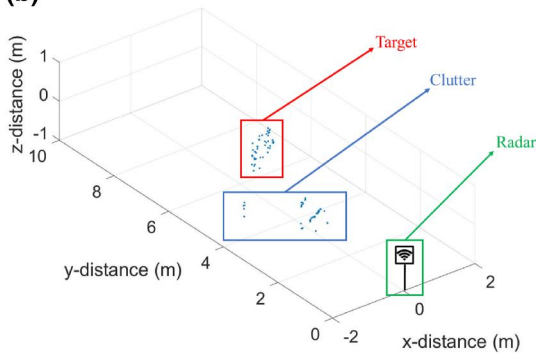


FIGURE 10 (a) Photograph of the experimental environment and (b) the target detection result in xyz coordinate system

describe a method for converting the point cloud data into images suitable for training our classifier.

The first step is to extract only the point cloud data corresponding to the desired target from the entire detection result. As shown in Figure 10, because signals reflected from surrounding objects and radar clutter are also detected, a process of cropping only detected points corresponding to the target is required. In this process, images received from the

camera mounted on the radar sensor, which is shown in Figure 6, were used as references. Based on this information, a virtual cuboid is created where the target exists, and only the point cloud data within the area is extracted. Then, all points outside the cube are considered clutter and are removed. Therefore, in consideration of the spatial size of pedestrians, cyclists and vehicles, only points corresponding to the desired target are extracted. Figure 11 shows examples of cropped point cloud data for pedestrians, cyclists, sedans and SUVs. As shown in the figure, unnecessary detected points are removed, and only information corresponding to the desired target remains.

Then, in the second step, the cropped point cloud data is orthogonally projected onto xy , yz and zx planes, as shown in Figure 12. Through this process, three radar detection results are obtained as if the target is viewed from the rear, side and top respectively. In other words, by looking at the target from various viewpoints, we can find out the spatial characteristics of the target. Unlike the conventional target classification methods that use the target detection result in one direction, we use the detection results in three directions for a single target.

The final step is to change the colour type of the image and resize it to improve the training efficiency. First, to increase the training speed, the size of the detection result for each target type is unified to 300×300 . In addition, images are usually represented as 3D data in red, green and blue (RGB) format. However, in the radar system, because information about the shape of the target is more important than colour information, we change the detection result to grey type to improve the training speed in the CNN. Even if the colour type is changed, the spatial characteristics of the target do not change. In addition, the training speed is improved because the size of the training data is reduced from 3D to 1D (i.e. only black information remains in RGB format). Through these three steps, the point cloud data generated from the radar sensor data is transformed into images that can be used for training the CNN-based classifier. Figure 13 is a flowchart summarising the step-by-step process of generating input images.

4.1.2 | Structure of proposed multi-view CNN-based classifier

Through the process described in Section 4.1.1, the point cloud data was converted into three images as if the target is viewed from three directions. Now, we describe the configuration of the data set for training the proposed classifier. In the process of training the classifier, 60%, 20% and 20% of the entire image data set are used as the training, validation and test data sets respectively. In other words, according to the above ratio, a total of 16,595 image data sets are divided into 9957, 3319 and 3319 images, respectively, as given in Table 3. In addition, 9957 images in the training data set consists of 2401 images of pedestrians, 1655 images of cyclists, 3019 images of sedans and 2882 images of SUVs. Moreover, both validation and test data sets have 3319 images, each consisting of 800

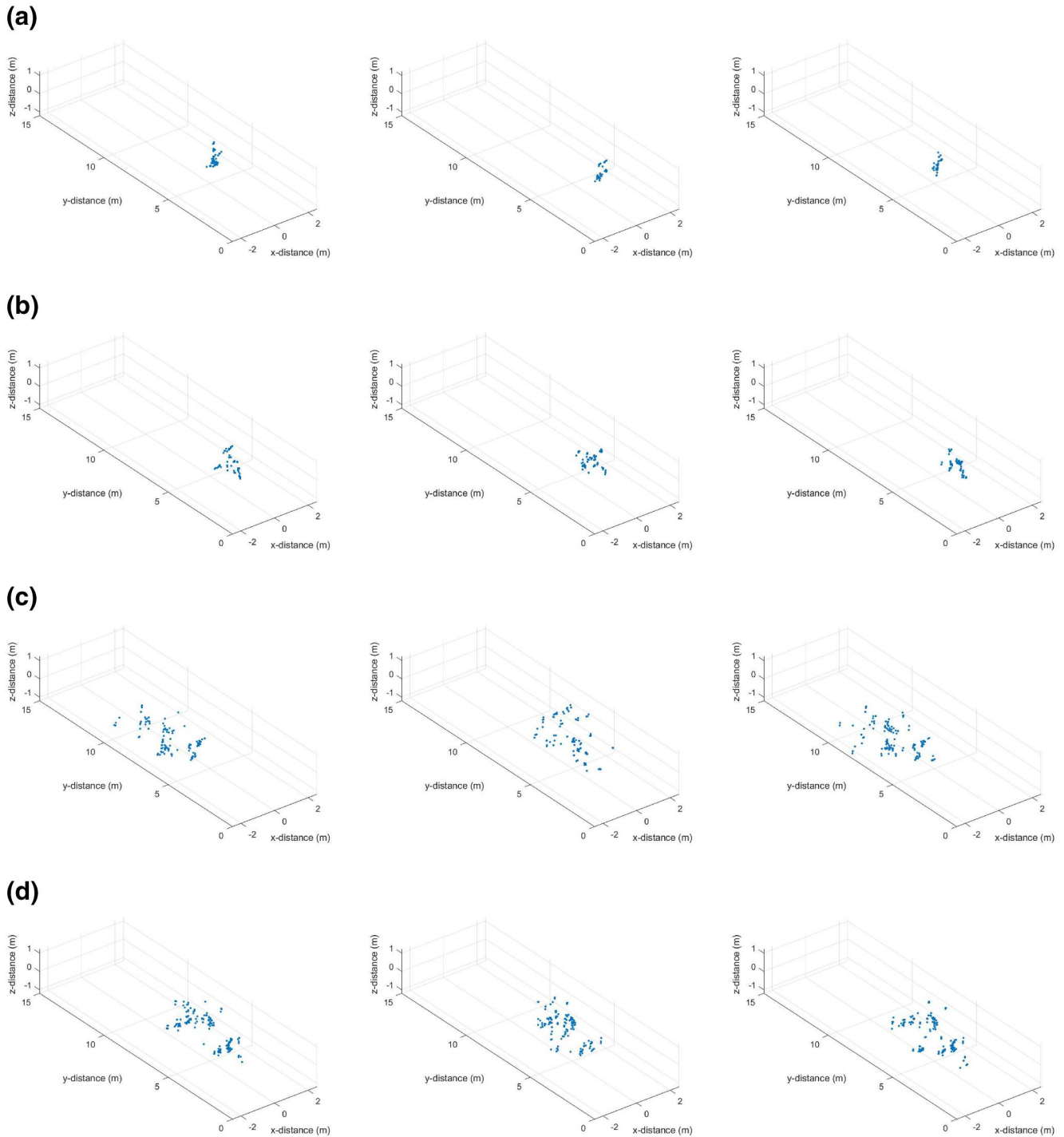


FIGURE 11 Cropped point cloud data: (a) for pedestrians, (b) cyclists, (c) sedans and (d) SUVs

images of pedestrians, 552 images of cyclists, 1006 images of sedans and 961 images of SUVs.

First, we designed a CNN structure that is trained using three images in parallel, which is shown in Figure 14. The proposed parallel-input CNN structure is a method for classifying target types by extracting target features from multiple viewpoints. As shown in the figure, the proposed structure consists of convolutional layers, normalisation layers, rectified linear unit (ReLU) layers, pooling layers and a

fully connected layer. The proposed parallel-input CNN structure classifies target detection images from three aspects into one of four target types (i.e. a pedestrian, a cyclist, a sedan or a SUV) through convolution operation. To determine the structure of the CNN-based classifier, the performance was evaluated while changing the number of hidden layers from 1 to 5. Each hidden layer consists of a convolutional layer, a batch normalisation layer, a ReLU layer and a max-pooling layer. In the training, the learning rate,

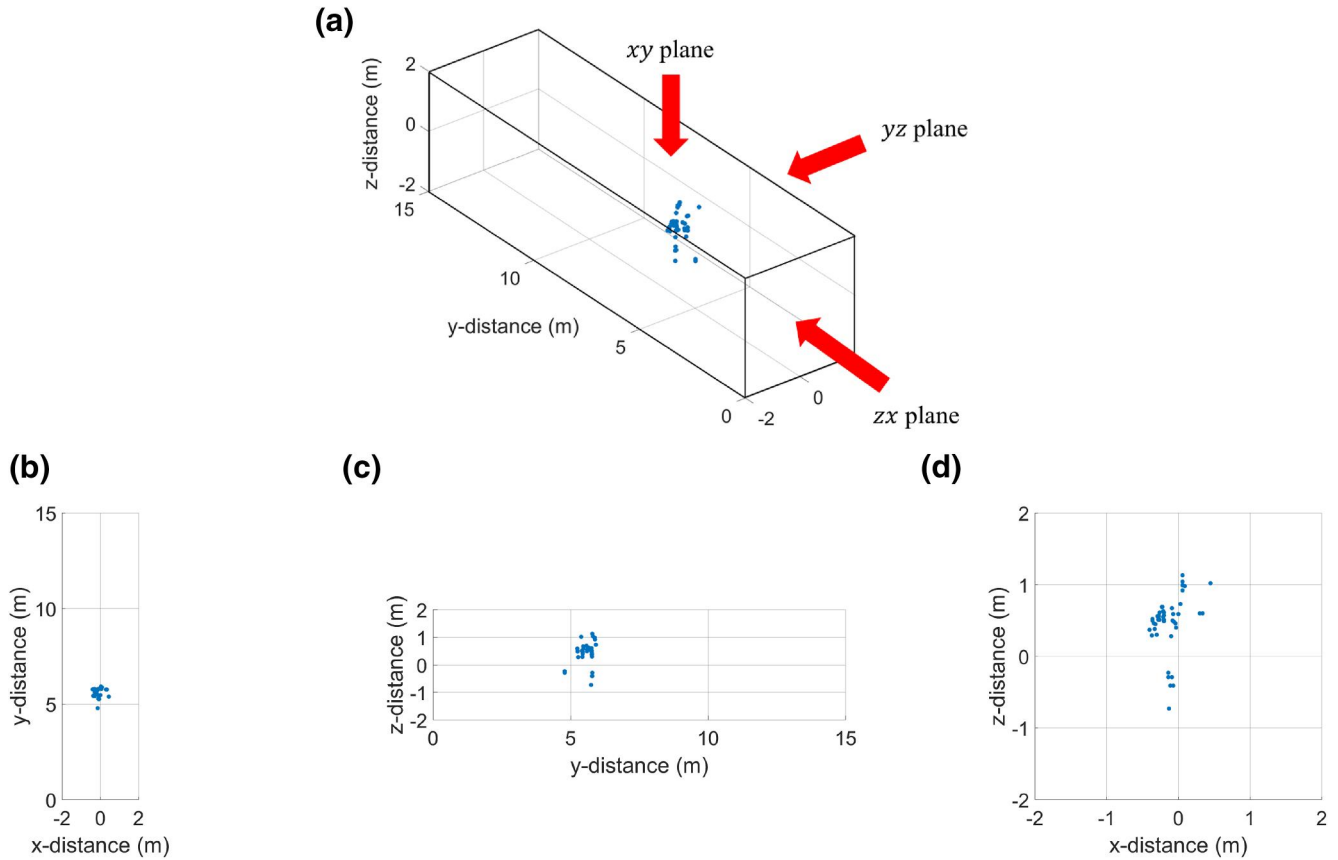


FIGURE 12 (a) Point cloud data for a pedestrian: (b) projection onto xy plane, (c) yz plane and (d) zx plane

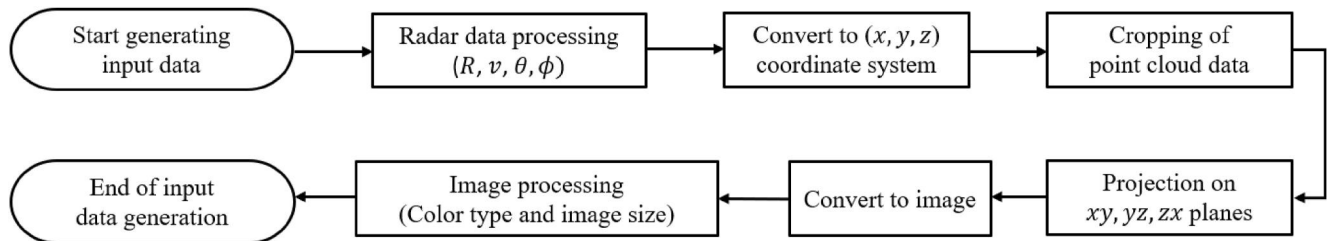


FIGURE 13 Flowchart for input data generation

Training data (60%)	Total data set (100%) validation data (20%)	Test data (20%)
2401 images of pedestrians	800 images of pedestrians	800 images of pedestrians
1655 images of cyclists	552 images of cyclists	552 images of cyclists
3019 images of sedans	1006 images of sedans	1006 images of sedans
2882 images of SUVs	961 images of SUVs	961 images of SUVs

TABLE 3 Configuration of the data set

which affects the accuracy and speed of training, was set to 0.001. In addition, we set the number of epochs to 6, where 1 epoch means the entire data set is used once for training. If the number of epochs is too small, underfitting occurs, and if the number of epochs is too large, overfitting occurs. Moreover, the process of shuffling the data was performed every epoch. Through the process of randomly selecting and

mixing data sets, the performance of the training model is accurately evaluated and overfitting is prevented.

Table 4 shows the average classification accuracy and training time according to the number of hidden layers. The training time was computed based on the AMD Ryzen 5 5600 CPU, GeForce RTX 2060 GPU and Samsung 16 GB RAM. As given in the table, the highest classification accuracy was

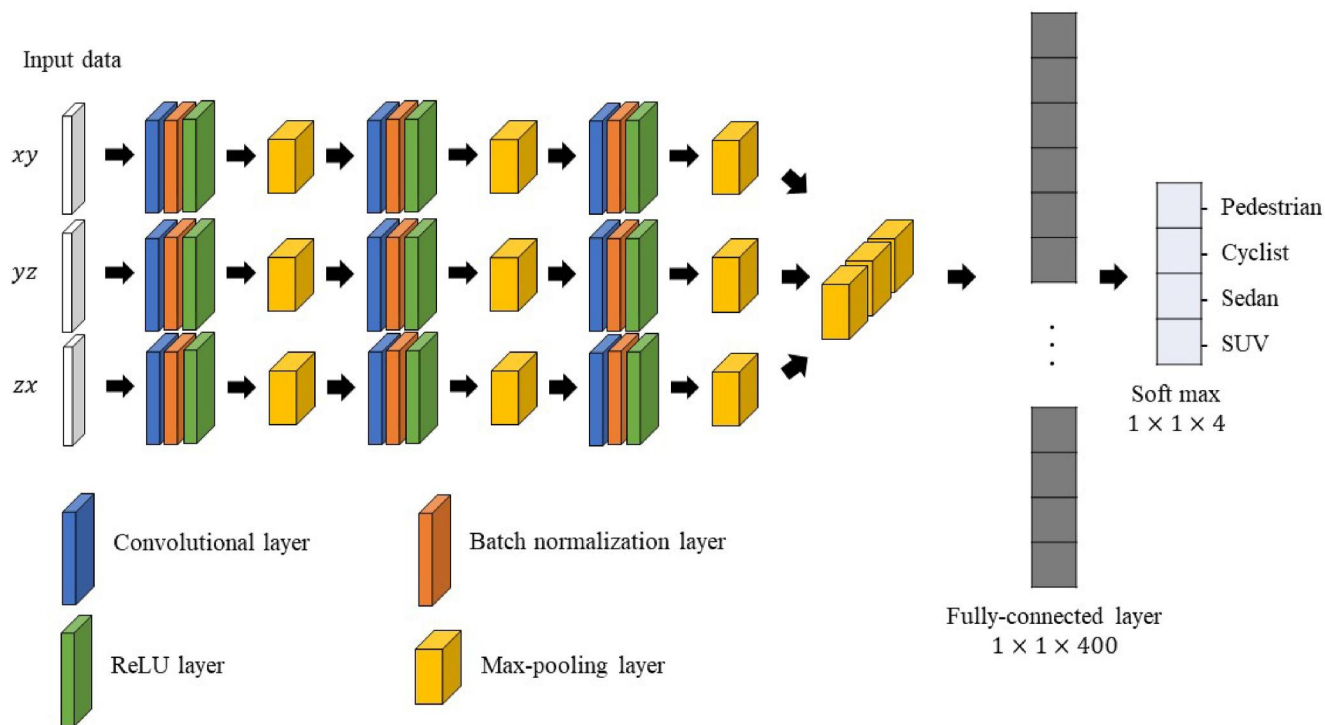


FIGURE 14 Proposed convolutional neural network (CNN) structure that is trained using three images in parallel

TABLE 4 Average classification accuracy and training time according to the number of hidden layers

The number of hidden layers	1	2	3	4	5
Average classification accuracy	98.92%	99.07%	99.55%	99.14%	98.70%
Average training time	5 m 40 s	5 m 48 s	8 m 18 s	9 m 58 s	12 m 53 s

achieved when three hidden layers were used. When there are more than three or more hidden layers, the training time increases slightly, but the classification accuracy does not increase significantly. Therefore, we decided to use a total of three hidden layers in our classifier.

In addition, we also designed a CNN classifier that takes three images in series, as shown in Figure 15. The serial-input CNN structure uses target detection images on the xy , yz and zx planes as 3D data, similar to the method using RGB format 3D data in general image classification. In summary, the main difference between serial-input CNN and parallel-input CNN structures is the way the three images for training are combined. In the serial-input CNN structure, three input images are merged and then training is performed, whereas in the parallel-input CNN structure, training is performed on each input image and then the training results are merged in the step just before the last fully connected layer. To evaluate only the difference in classification performance due to structural differences, the values of all variables related to network training were set identically in both serial-input and parallel-input CNN structures.

4.2 | Performance evaluation

First, we evaluated the classification performance of the parallel-input CNN structure. As shown in Figure 14, input

images are classified into one of four types by the proposed parallel-input CNN structure. Table 5 shows the confusion matrix for the parallel-input CNN structure. The proposed structure classified four types of targets with an average classification accuracy of 99.13% or higher (As given in Table 3, because a different number of input images were used for each target type, the average classification accuracy is slightly different from calculating the average of the components on the diagonal of the confusion matrix.). As shown in the table, the classification accuracy for cyclists was relatively low compared to the classification accuracies for pedestrians, sedans, and SUVs. Because the spatial size of cyclists is similar to that of pedestrians rather than that of vehicles, cyclists tend to be classified as pedestrians. In addition, Table 6 shows the confusion matrix for the serial-input CNN structure and the four types of targets were classified with an average accuracy of 99.16% or more. Similar to the case of the parallel-input CNN structure, the class with the lowest classification accuracy in the serial-input CNN structure was the cyclist, which was 97.60%. As mentioned before, cyclists tend to be classified as pedestrians in both structures because of their spatial size.

Finally, the classification performance of the proposed method was compared with the performance when only one of the xy , yz and zx plane target detection results was used. In

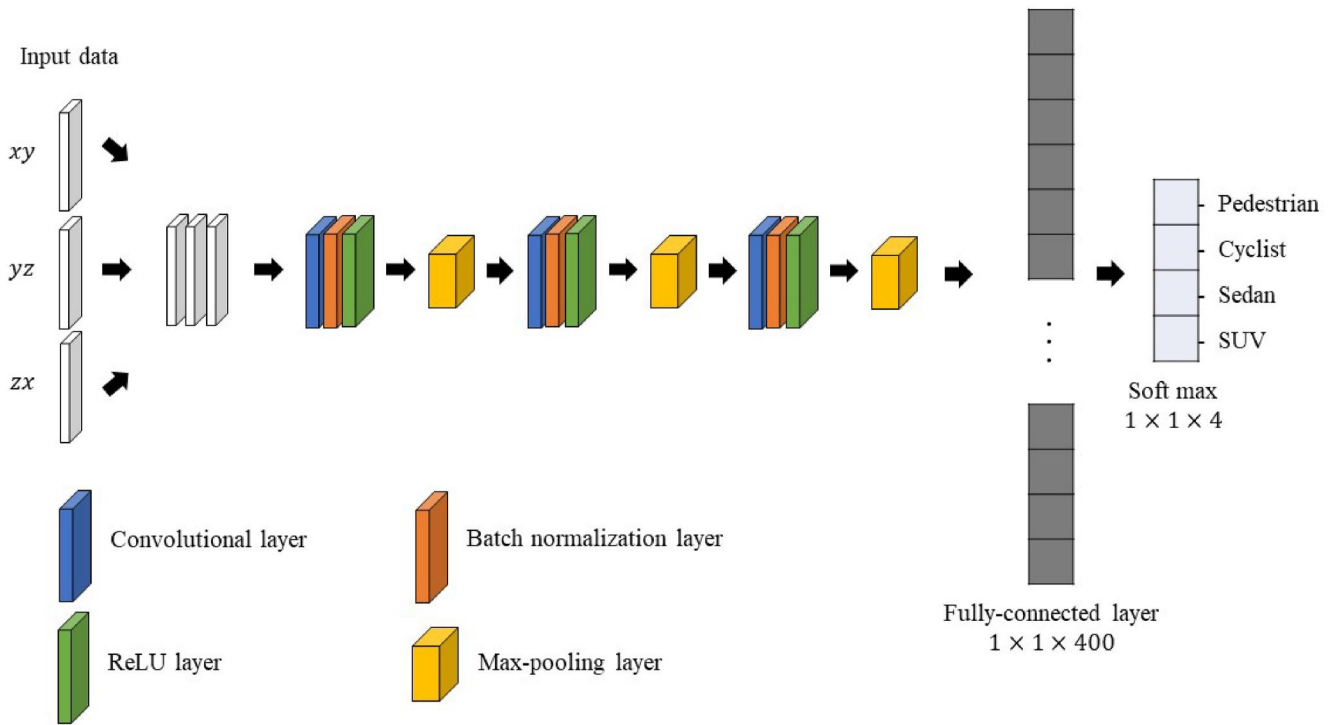


FIGURE 15 Proposed convolutional neural network (CNN) structure that is trained using three images in series

TABLE 5 Confusion matrix in the parallel-input convolutional neural network (CNN) structure

Actual class	Estimated class			
	Pedestrian	Cyclist	Sedan	SUV
Pedestrian	98.18%	1.82%	0%	0%
Cyclist	2.21%	97.61%	0.18%	0%
Sedan	0%	0%	99.90%	0.10%
SUV	0%	0%	0%	100.00%

TABLE 6 Confusion matrix in the serial-input convolutional neural network (CNN) structure

Actual class	Estimated class			
	Pedestrian	Cyclist	Sedan	SUV
Pedestrian	98.45%	1.55%	0%	0%
Cyclist	2.40%	97.60%	0%	0%
Sedan	0.10%	0%	99.80%	0.10%
SUV	0%	0%	0.10%	99.90%

other words, except for the part that synthesises the input data in Figure 15, a classifier with the same structure was trained using only a single plane image. Table 7 summarises the time required for training, the classification accuracies for the training and test sets, and the class with the lowest classification accuracy.

When comparing the proposed parallel-input and serial-input CNN structures, the former takes much more time to

train than the latter. This is because training is performed on each input image and then merged in the step just before the last fully connected layer in the parallel-input CNN structure. For the training data set, the average classification accuracy was 99.13% for the parallel-input CNN structure and it was 99.16% for the serial-input CNN structure. In addition, the average classification accuracy for the test data set was 99.30% for the parallel-input CNN structure and it was 98.92% for the serial-input CNN structure.

When training was performed with only one detection result in the xy , yz and zx planes, the training time was reduced by 60% compared to the proposed method of merging three detection results. However, the average classification accuracies for the training and test sets were lowered. The proposed parallel-input and serial-input CNN structures use all three spatial characteristics of length, height and width when classifying the target. However, only two of these characteristics are used when classifying the target using only the detection result in a single plane. For example, because cyclists and pedestrians have similar heights and widths, length information is required to increase classification accuracy. However, when only the detection result in the zx -plane is used, there is no information in the y -axis direction indicating the length characteristic. Thus, cyclists are misclassified as pedestrians and vice versa. Also, the advantage of using all three detection results is shown in the class with the lowest classification accuracy. As shown in the table, the class with the lowest classification accuracy in the proposed method is over 97.6%. However, when only the detection result of a single plane is used, the classification accuracy drops to 93.1% on average, which is 4.5%*p* lower

TABLE 7 Performance comparison with convolutional neural network (CNN) structures using a single input

Input data	xy plane	yz plane	zx plane	Images merged in parallel	Images merged in serial
Time required for training	2 m 5 s	2 m 25 s	2 m 7 s	7 m 16 s	4 m 20 s
Classification accuracy for training set	97.8%	97.95%	96.11%	99.13%	99.16%
Classification accuracy for test set	98.00%	98.80%	96.20%	99.30%	98.92%
Class with the lowest classification accuracy	94.63 (pedestrian)	95.03 (cyclist)	89.7% (cyclist)	97.61% (cyclist)	97.60% (cyclist)

TABLE 8 Performance comparison with GoogLeNet [20], ResNet [21] and SqueezeNet [22]

Classifier	GoogLeNet [20]	ResNet [21]	SqueezeNet [22]
Time required for training	35 m 3 s	19 m 18 s	10 m 17 s
Classification accuracy for training set	98.82%	99.70%	95.51%
Classification accuracy for test set	99.1%	99.5%	95.1%
Class with the lowest classification accuracy	96.34% (cyclist)	98.89% (cyclist)	87.7% (cyclist)

TABLE 9 Confusion matrix in the PointNet [23]

Actual class	Estimated class			
	Pedestrian	Cyclist	Sedan	SUV
Pedestrian	84.76%	14.21%	0.69%	0.34%
Cyclist	16.43%	82.37%	1.20%	0%
Sedan	0.38%	4.59%	94.78%	0.25%
SUV	0.31%	2.54%	1.73%	95.42%

than the accuracy of the proposed method. In all methods, the time taken to classify a new input image was approximately the same.

Finally, with the same point cloud data set, we compared the classification performance with the well-known deep learning techniques, such as the GoogLeNet [20], ResNet [21], SqueezeNet [22] and PointNet [23]. First, we compared the classification performance with the GoogLeNet, ResNet and SqueezeNet, which are widely used for image classification. The time required for training, the classification accuracies for the training and test sets and the class with the lowest classification accuracy are given in Table 8. Among the three CNN-based structures, the ResNet showed the highest classification accuracy of 99.70%. However, compared to the proposed serial-input CNN structure, the training time increased by 445%. Therefore, although the two methods are similar in terms of classification performance, our proposed method is much more efficient in terms of training time. Moreover, the confusion matrix when the PointNet was applied is given in Table 9. As shown in the table, the average classification accuracy was 89.33%, which was 9.8%*p* lower than that of the proposed method. In particular, the classification accuracy for cyclist was the lowest at 82.37%. In general, a large number of points are required for the PointNet to accurately classify a target. If the range and angle resolution of the radar system is further improved, the classification performance of the PointNet is expected to improve.

5 | CONCLUSION

In this paper, we proposed the multi-view CNN-based target classification method for the high-resolution automotive radar system. First, we processed the sensor data acquired from the MIMO FMCW radar sensor and obtained point cloud data for four types of targets (i.e. pedestrians, cyclists, sedans and SUVs). To find the spatial characteristics of the target, the point cloud data in the 3D space was orthogonally projected onto the *xy*, *yz* and *zx* planes. Then, the CNN-based classifiers that synthesise three images in series and parallel were proposed. Finally, the classification performance and training time of the proposed CNN structures were evaluated. The proposed classification method showed 4.5%*p* higher average classification accuracy than the classifier using the detection result of one plane in terms of the target with the lowest classification accuracy. In addition, the proposed method showed better classification performance than the well-known deep learning-based image classification algorithms and was more efficient in terms of training time.

ACKNOWLEDGEMENTS

Seungheon Kwak and Hangeol Kim contributed equally to this work. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00,237). The authors thank bitsensing Inc., Seongnam-si, Gyeonggi-do, Republic of Korea, for providing the high-resolution radar sensor and giving access to the raw radar sensor data.

CONFLICT OF INTEREST

There is no conflict of interest.

DATA AVAILABILITY STATEMENT

Author elects to not share data.

ORCID

Seongwook Lee  <https://orcid.org/0000-0001-9115-4897>

REFERENCES

1. Rohling, H., Möller, C.: Radar waveform for automotive radar systems and applications. In: 2008 IEEE Radar Conference, pp. 1–4 (2008)
2. MIMO Radar. <https://www.ti.com/lit/an/swra554a/swra554a.pdf>. Accessed 15 April 2022
3. Meyer, M., Kuschik, G.: Automotive radar dataset for deep learning based 3D object detection. In: 2019 16th European Radar Conference (EuRAD), pp. 129–132 (2019)
4. Schumann, O., et al.: RadarScenes: A Real-World Radar Point Cloud Data Set for Automotive Applications, pp. 1–8 (2021). arXiv:2104.02493v1
5. Palfy, A., et al.: Multi-class road user detection with 3+1D radar in the view-of-Delft dataset. *IEEE Rob. Autom. Lett.* 7(2), 4961–4968 (2022)
6. Zheng, L., et al.: Tj4DRadSet: A 4D Radar Dataset for Autonomous Driving. arXiv:2204.13483v2, pp. 1–6 (2022)
7. Lee, S., et al.: Human-vehicle classification using feature-based SVM in 77-GHz automotive FMCW radar. *IET Radar Sonar Navig.* 11(10), 1589–1596 (2017)
8. Heuel, S., Rohling, H.: Pedestrian classification in automotive radar systems. In: 2012 13th International Radar Symposium, pp. 39–44 (2012)
9. Zhao, Z., et al.: Point cloud features-based kernel SVM for human-vehicle classification in millimeter wave radar. *IEEE Access* 8, 2169–3536 (2020)
10. Kim, Y., Alnujaim, I., Oh, D.: Human activity classification based on point clouds measured by millimeter wave MIMO radar with deep recurrent neural networks. *IEEE Sensor. J.* 21(12), 13522–13529 (2021)
11. Cheng, Y., Liu, Y.: Person reidentification based on automotive radar point clouds. *IEEE Trans. Geosci. Rem. Sens.* 60, 1–13 (2022)
12. Bai, J., et al.: Radar transformer: an object classification network based on 4D MMW imaging radar. *Sensors* 21(11), 1–16 (2021)
13. Bai, J., et al.: Traffic participants classification based on 3D radio detection and ranging point clouds. *IET Radar Sonar Navig.* 16(2), 278–290 (2022)
14. Winkler, V.: Range Doppler detection for automotive FMCW radars. In: 2007 European Radar Conference, pp. 166–169 (2007)
15. Patole, S.M., et al.: Automotive radars: a review of signal processing techniques. *IEEE Signal Process. Mag.* 34(2), 22–35 (2017)
16. Gross, F.B.: *Smart Antennas for Wireless Communications with MATLAB*. McGraw-Hill (2005)
17. Krim, H., Viberg, M.: Two decades of array signal processing research: the parametric approach. *IEEE Signal Process. Mag.* 13(4), 67–94 (1996)
18. AIR 4D. <https://bitsensing.com/ko/air-4d/>. Accessed 15 April 2022
19. Cohen, M.N.: An overview of high range resolution radar techniques. In: NTC '91 – National Telesystems Conference Proceedings, pp. 107–115 (1991)
20. Szegedy, C., et al.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015)
21. He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
22. Iandola, F.N., et al.: SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and <0.5MB Model Size. arXiv:1602.07360v4, pp. 1–13 (2016)
23. Qi, C.R., et al.: PointNet: deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 77–85 (2017)

How to cite this article: Kwak, S., et al.: Multi-view convolutional neural network-based target classification in high-resolution automotive radar sensor. *IET Radar Sonar Navig.* 17(1), 15–26 (2023). <https://doi.org/10.1049/rsn2.12320>